EXPERIMENTS ON LANGUAGE ERRORS IN AVIATION MAINTENANCE: ASIA

Colin G. Drury and Jiao Ma
University at Buffalo, Department of Industrial Engineering
438 Bell Hall, Buffalo, NY 14260
drury@buffalo.edu

The Federal Aviation Administration has raised many issues concerning the outsourcing of maintenance to foreign repair stations and recommends establishing a method for determining whether language barriers result in maintenance deficiencies. This work addresses concerns that non-native English speakers may be prone to an increased error rate that could potentially affect airworthiness. This paper presents Year 2 of the project. We used the seven scenarios of language error developed in Year 1 as the basis for our data collection effort to quantify the frequency of error. An intervention experiment has been designed and tested using a sample of 200 maintenance personnel from countries in Asia. The interventions were found to increase document comprehension performance. Participants tended to maintain a constant accuracy level, with performance changes coming mainly from speed differences.

INTRODUCTION

In 2001, the Federal Aviation Administration raised many issues concerning the outsourcing of maintenance to foreign repair stations in considering changes to domestic and foreign Federal Air Regulations, recommending that:

> *"The FAA should establish a method for determining whether language barriers result in maintenance deficiencies."*

This project is a direct response to these concerns that non-native English speakers, in repair stations in the USA and abroad, may be prone to an increased error rate that could potentially affect airworthiness. The documentation for repair provided by an English speaking airline is always in English, and this documentation must be used to govern all maintenance tasks, despite a potentially large proportion of mechanics who do not use English as a native language. This paper follows our 2003 HFES paper (Drury and Ma, 2003) and describes the first experiment using a methodology for quantifying the effectiveness of possible countermeasures to language errors.

As noted in our 2003 paper, this project developed seven scenarios of language error based on visits to sites in the USA and the UK; it also provided a model for these unique communication errors based on the communications literature and an analysis of several databases (e.g., NASA/ASRS). Many references to communication theories and studies of outsourcing were given in Drury and Ma (2003) and will not be repeated here.

The seven scenarios found were:

Scenario 1: "The Mechanic (Aircraft Maintenance Technican, AMT) or Inspector was not able to communicate verbally to the level required for adequate performance."

Scenario 2: "The Mechanic (AMT) or Inspector and the person to whom they were speaking did not realize that the other had limited English ability."

Scenario 3: "Native English speakers with different regional accents did not understand each others' communications."

Scenario 4: "The Mechanic (AMT) or Inspector did not understand a safety announcement over the Public Address (PA) system."

Scenario 5: "The Mechanic (AMT) or Inspector did not fully understand a safety placard."

Scenario 6: "The Mechanic (AMT) or Inspector did not fully understand documentation in English, for example a Work Card or a Manual."

Scenario 7: "The Mechanic (AMT) or Inspector did not fully understand a document translated from another language into their native language."

In our work, we have been visiting sites worldwide to measure the frequency of these scenarios, and evaluating the effectiveness of countermeasures.

A survey conducted by a major manufacturer showed that language skill varied (as expected) by world region, and that not all sites with lower language skills translated documents into the native language. Our analysis of the survey data reported earlier found that two strategies used to reduce the potential for language errors were (a) translation into the native language, and (b) conducting face-to-face meetings in the native language. However, only about 17% of airlines in the region that most often used translation (Asia) actually translated maintenance documents into the native languages. Even among the group of 8 airlines who reported the lowest English speaking ability, only 2 modified the English documents in any way. Other strategies of intervention found in our site visits included

having a bilingual English/native language speaker (e.g., lead, engineer) assist the mechanic with the English documentation, and/or providing a glossary of key words between the native language and English. Finally, our own earlier research into the artificial maintenance language called European Association of Aerospace Industries (AECMA) Simplified English (e.g., Chervak, Drury and Ouellette, 1996) had shown it to be an effective error reduction technique, particularly for non-native English speakers and for complex work documents.

Thus, we planned to compare four potential language error reduction interventions:

- The translation of a document into AECMA Simplified English
- The provision of a Glossary
- The provision of a bilingual coach
- The translation of a document and all related materials into a native language

Some of these methods can be combined, for example the provision of both a Glossary and a bilingual coach, or the addition of AECMA Simplified English to all conditions except for translation into the native language. Finally, for comparison, a baseline condition, no intervention, was required. This paper describes briefly the first two experiments conducted within this framework, and the main data collection in one region, Asia.

METHODOLOGY

Measures

To test for how potential documentation errors can be reduced, we measured the effectiveness of document comprehension. In the study, a single task card was given to participants with a 10-item questionnaire to test comprehension. The methodology was validated in our previous research (e.g., Chervak, et al., 1996; Drury, Wenner and Kritkausky, 1999). The comprehension score was measured by the number of correct responses, with time taken to complete the questionnaire as an additional measure.

Task Cards

We selected two task cards, one "easy" and one "difficult," from four task cards used in the previous research, because it had already been found that task difficulty affected the effectiveness of one strategy, Simplified English. As was expected, the use of Simplified English had a larger effect on more complex task cards (Chervak and Drury, 2003). The complexity of these task cards was evaluated by Boeing computational linguists and University of Washington technical communications researchers considering word count, words per sentence, percentage passive voice, and the Flesch-Kincaid reading score. The cards differed on all measures.

Both of the task cards were then prepared in the AECMA Simplified English versions, which were also critiqued by experts from Boeing, the University of Washington, and the American Institute of Aeronautics and Astronautics (AIAA) Simplified English Committee.

Pre-Test Design

First, to test the design and materials, two pilot studies were conducted, one using 15 English-speaking maintenance personnel from sites in the USA and the UK, and the other using 40 Native Chinese speaking engineering graduate students at the University at Buffalo, SUNY. These tests successfully proved the evaluation methodology, and eliminated one condition (glossary plus bilingual coach) as participants did not make use of both. Full details were given in our 2004 HFES paper (Drury and Ma, 2004).

Design

A three-factor design was used with participants fully nested under all conditions:

Task card Complexity: 2 levels - Simple
- Complex

Task card Language: 2 levels - Simplified English
- Not Simplified English

Language Interaction: 4 levels - No intervention (English)
- English with glossary
- English with coach
- Full Chinese translation

Choice of Participants and Sites

There are several reasons to collect data from MROs located in Asia, especially China. First, in our analysis of the manufacturer's survey data, we found that about 30% of users in Asia had a very limited English speaking ability, another 40% were able to conduct simple conversations; about 40% of the users were able to work effectively with only written maintenance/inspection related documents, and another 15% had very little English reading ability. Compared with North America and Europe, Asia has a much smaller base of English-using mechanics. Second, the Asia-Pacific region is poised to be one of the strongest growth engines for the foreseeable future for the maintenance, repair and overhaul industry (*Overhaul & Maintenance*, 2002). U.S. and European airlines continue to ship wide-body aircraft to East Asia to take advantage of low labor costs. Almost half of the top 10 Asian MROs are located in China. According to *Aviation Week & Space Technology*, "the Civil Aviation Administration of China (CAAC) is confident that despite the downturn in the global

airline industry, more maintenance, repair and overhaul (MRO) joint venture companies will be set up with Chinese airlines within the next two years" (Dennis, 2002).

Participants were tested individually or in small groups. After obtaining Informed Consent and completing demographic questions, the participants were given one of the four task cards and its associated comprehension questions. They were timed, but instructions emphasized accuracy. After the completion of the comprehension task, the participants were given the Accuracy Level Test (Carver, 1987), for the required 10 minutes to act as a potential covariate in our analysis. This test used a total of 100 words with a forced synonym choice among three alternatives, and produced on the scale of reading grade level. It has been validated against more detailed measures of reading level (Chervak, Drury, Ouellette, 1996).

Preparation of the Data Collection Packet for Asia

The translation process took place in two steps. A native Chinese research assistant (9 years as an engineering major), who is very familiar with the task cards, took a lead in translating the packet. A large number of technical and language references were consulted. The principal investigator and other domain experts (e.g., native Chinese mechanical engineers in the Department of Aerospace and Mechanical Engineering at the University at Buffalo, SUNY) were consulted on the technical details (e.g., lockwire). Then both translated task cards, and original packets of data collection material were submitted to a retired professor from the Department of Avionics, Civil Aviation University of China (CAUC) for review.

We developed an English/Chinese glossary for each task card. We had two native English speaking engineering graduate students and two native Chinese speaking engineering graduate students read through all the task cards and circle all the words/phrases/sentences they did not comprehend, or even those about which they were slightly unsure. We built up this glossary to be as comprehensive as possible, including nouns, verbs, adjectives, abbreviations, etc.

Results from Asia: Intervention Performance

This test used 200 participants from six sites in mainland China and Hong Kong First, in contrast to the pre-tests, there was almost no negative correlation between accuracy and time for the comprehension test ($r = -0.210$, $p = 0.09$). There were moderate correlations of both with Years as an AMT ($p = 0.061$ and 0.008) and Years Learning English ($p = 0.005$ and 0.006). A third measure was created by dividing Accuracy by Time to give a combined overall Performance score.

Reading Level was tested as a covariate, but was not significant in any of three GLM ANOVAs of Accuracy, Time, and Performance. Years as an AMT was a significant covariate in all three measures, but did not change the significance pattern of the three factors, so results of ANOVAs rather than ANCOVAs will be presented here. The surprising overall result was that Accuracy of comprehension did not vary with any of the factors except Site which was included as a main effect only ($F(5, 179) = 2.58$, $p < 0.028$).

The sites were different on Time and Performance measures ($F(5,177) = 7.88$, $p < 0.001$, and $F(5, 177) = 5.46$, $p < 0.001$), with the two sites in Hong Kong being more rapid

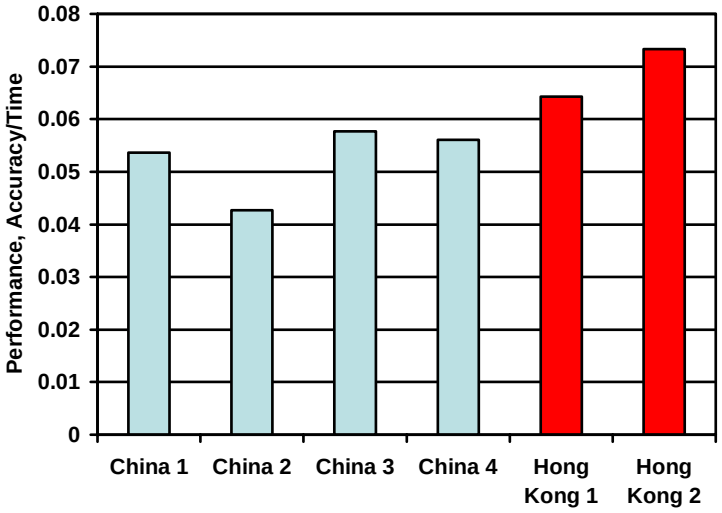

Figure 1: Performance Results by Site

and having a higher performance than the mainland China sites (Figure 1).

The other significant main effects were for Intervention and Task Card. Intervention was significant for Time ($F(3,179) = 7.57$, $p < 0.001$) and Performance ($F(3,179) = 2.99$), while Task Card was significant at ($F(1,179) = 15.43$, $p < 0.001$) and ($F(1,179) = 5.02$, $p = 0.026$) respectively. The Easy task card had a performance score of 0.058 while the Difficult task card scored 0.052. Interventions grouped into two sets, with all three active interventions faster than the baseline condition. In terms of Performance, the comparisons are shown in Figure 2. Note that the use of AECMA Simplified English had no significant effect on any measures. Also, no interactions among any factors reached significance, simplifying the interpretation of results.

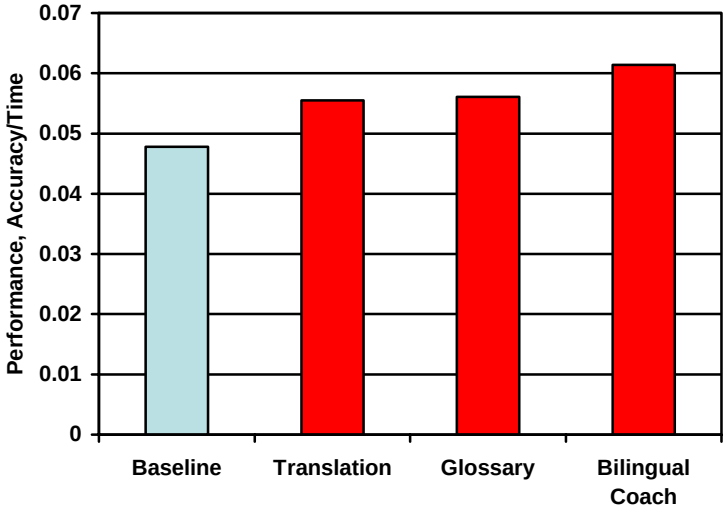

Figure 2: Performance comparisons by Intervention.

Results from Asia: Scenario Analysis

In addition to the evaluation of the interventions, we used a questionnaire to determine the relative incidence of the seven scenarios developed earlier. A number of measures of incidence were used, including estimates of the time since last occurrence. Here we present only the overall response to "Have you ever encountered an error of this type?" The incidence of each scenario is shown in Figure 3 for mainland China and Hong Kong separately. Note that the four most frequently encountered scenarios (1, 2, 6 and 7) are concerned with directly work related verbal and written ability. The other three scenarios concern regional accents, and less-work-related events. The misunderstood translations (Scenario 7) often referred to translations from English by aircraft manufacturers for whom English is not the native language.

For the response to factors most associated with these scenarios, GLM ANOVA of the percentage encountering each incident by Factor was performed, with Region and Scenario as additional independent variables. All main effects and interactions except Factor × Country were significant at $p < 0.01$ or better. The responses divided into two groups, one seen as highly related to the incident and one less related. These are:

Highest Related to Scenarios

Task is Complex
Task Instruction is complex
AMT's inadequate written English
AMT's inadequate verbal English
Time pressure on AMT

Lowest Related to Scenarios

Poor communication equipment
AMT does not ask for help
AMT uses native language under stress
Unwilling to expose lack of English

A similar analysis was performed for the ten factors potentially mitigating language errors. The GLM ANOVA gave significance at $p < 0.01$ for Factor, Region, and their interaction. As with causal factors, the results grouped into two:

Highest Related to Scenarios

Translated documents
Consistent terminology
Document uses good design practices
Use of aircraft for communication
AMT is familiar with the job

Lowest Related to Scenarios

AMT has passed comprehension test
AMT is certified for that job
Translator is available to AMT
Jobs is assigned based on English ability
AMT team with English speaker

As with causal factors, the highest group included the physical changes, plus in this case job familiarity. The lowest group was mainly individual and social interventions.

Finally, an analysis of how errors are discovered was performed. Only Scenario, Factor, and the Factor × Country were significant (at $p < 0.02$). Again, there was a grouping of the Factors, this time into 3 groups:

Highest Related to Scenarios

AMT asked for assistance/clarification

Medium Related to Scenarios

AMT appeared perplexed
Resulting physical error was detected

Lowest Related to Scenarios

AMT agreed with everything said
AMT did not understand at buy-back
AMT closed access prematurely

From these groupings, note that the least commonly found were either an unusual behavior, or events later in the maintenance/inspection process.

CONCLUSIONS

On our site visits, we conducted two main studies. The first was a direct test of the effectiveness of four interventions and the second an evaluation of the incidence and causal factors in seven previously-developed language errors scenarios.

The interventions experiment used a baseline condition of English documents, and then added translation

(including the test form), a glossary, a bilingual coach, and a combination of these last two conditions. We used two levels of task card difficulty, each with and without Simplified English. This made a three-factor factorial experiment (Intervention × Difficulty × Simplified English), with various covariates.

On the samples tested so far, the results are encouraging. While there were some differences between regions, differences between interventions were consistent across regions. All of the interventions had some effect, although mainly on the times and our performance measure, rather than on accuracy *per se*. If this indeed reflects practice, then maintenance personnel appear to slow down when they find language difficult, rather than making more errors at a constant speed.

The analysis of the incidence and factors data suggest that most of the causal factors in language errors are seen to be either directly document-related or time pressures. The factors least related are much more behavioral or communications channel related. A similar result was found for mitigating factors. These findings give some credence to our use of the documentation interventions, which should address four of the five highest related causal factors.

Our next task is to repeat this experiment in other continents. The current plan is to visit locations in Central and South America in Fall 2004 and Europe in Spring 2005.

REFERENCES

Carver, R. P. (1987). *Technical Manual for the Accuracy Level Test*. REVTAC Publications, Inc.

Chervak, S., and Drury, C. G. (2003). Effects of Job Instruction on Maintenance Task Performance. *Occupational Ergonomics*. Vol.3, Issue 2, 121-131

Chervak, S., Drury, C. G., and Ouellette, J. L. (1996). Simplified English for Aircraft Workcards. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, 303-307.

Dennis, W. (2002). MRO to Grow in China. *Aviation Week and Space Technology*.

Drury, C. G. and Ma, J. (2003). Do Language Barriers Result in Aviation Maintenance Errors? *Human Factors and Ergonomics Society 47th Annual Meeting Proceedings*, Denver, Colorado, October 13-17, 2003.

Drury, C. G. and Ma, J. (2004). Experiments on Language Errors in Aviation Maintenance *Human Factors and Ergonomics Society 47th Annual Meeting Proceedings*, New Orleans LA September 20-24, 2004.

Drury, C. G. and Ma, J. (2003). *Language Errors in Aviation Maintenance: Year 1 Interim Report*, Reports to William J. Hughes Technical Center, the Federal Aviation Administration under research grant #2002-G-025.

Drury, C.G., Wenner, C., and Kritkausky, K. (2000). Information Design Issues in Repair Stations. *Proceedings of Tenth International Symposium on Aviation Psychology*. Columbus, OH.

MRO in Asia-Pacific Region (2002). Aviation Week's ShowNews Online: www.aviationnow.com. Online resources: *http://www.awgnet.com/shownews/02asia1/mro09.htm* & *Overhaul & Maintenance*.

Patel, S., Drury, C. G.and Lofgren, J. (1994). Design of Workcards for Aircraft Inspection. *Applied Ergonomics*, **25(5)**, 283-293.

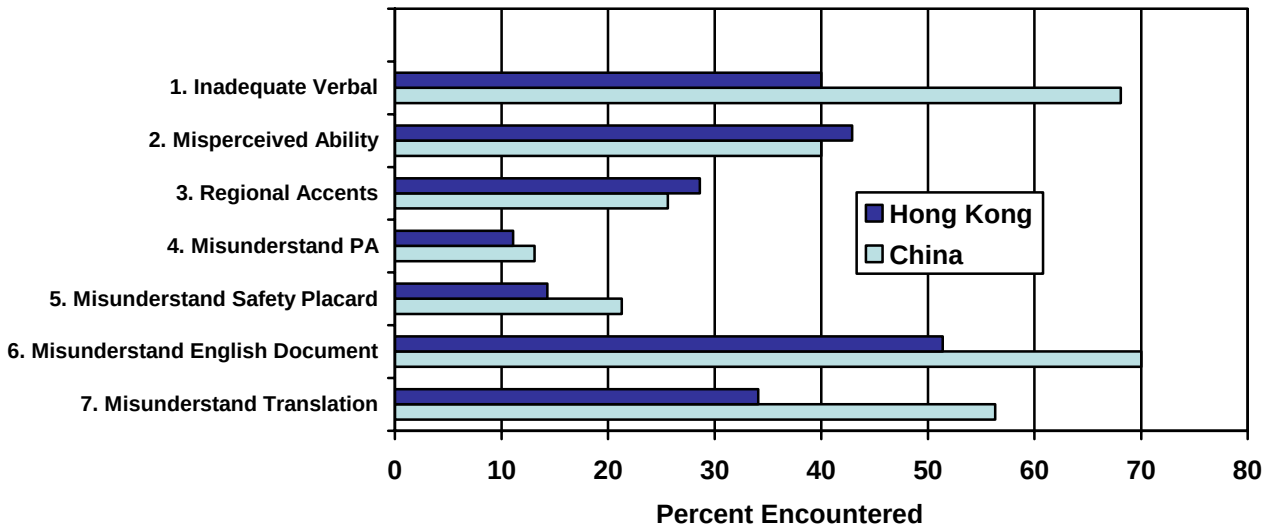

Figure 3: Relative frequency with which each of the seven scenarios was encountered.